DOCUMENT RESUME

ED 331 877 TM 016 426

AUTHOR Mazor, Kathleen M.; And Others
TITLE The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic.
PUB DATE Apr 91
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; *Item Bias; Item Response Theory; *Sample Size; *Simulation; Test Items
IDENTIFIERS *Mantel Haenszel Procedure

ABSTRACT

The Mantel-Haenszel (MH) procedure has become one of the most popular procedures for detecting differential item functioning. Valid results with relatively small numbers of examinees represent one of the advantages typically attributed to this procedure. In this study, examinee item responses were simulated to contain differentially functioning items, and then were analyzed at five sample sizes (2,000, 1,000, 500, 200, and 100) to compare detection rates. Five different 75-item tests were generated for each group. Results show that the MH procedure missed more than 30% of the differentially functioning items when groups of 2,000 were used. When 500 or fewer examinees were retained in each group, more than 50% of the differentially functioning items were missed. The items most likely to be undetected were those that were difficult, those with a small difference in item difficulty between the two groups, and poorly discriminating items. Three tables and four graphs describe the simulations. (Author/SLD)

The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic

Kathleen M. Mazor, Brian E. Clauser, Ronald K. Hambleton
University of Massachusetts at Amherst

Abstract

The Mantel-Haenszel (MH) procedure has become one of the most popular procedures for detecting differential item functioning. Valid results with relatively small numbers of examinees is one of the advantages typically attributed to this procedure. In this study, examinee item responses were simulated to contain differentially functioning items, and then were analyzed at five sample sizes to compare detection rates. Results showed the MH procedure missed 25 to 30 percent of the differentially functioning items when groups of 2000 were used. When 500 or fewer examinees were retained in each group more than 50 percent of the differentially functioning items were missed. The items most likely to be undetected were those which were most difficult, those with a small difference in item difficulty between the two groups, and poorly discriminating items.

Lab Report 211

The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic[1,2]

Kathleen M. Mazor, Brian E. Clauser, Ronald K. Hambleton

Standardized tests have become an integral part of modern society. Test results are often used to inform decisions regarding placement, advancement and competency. Because of the importance of such decisions those involved in the testing process are concerned that the results of such tests be valid. Differential item functioning is one threat to test validity.

Differential item functioning (DIF) is said to be present when examinees of the same ability, but of different groups, have differing probabilities of getting an item correct. Thus, one group is at a relative advantage. Often the group which has the advantage on a DIF item is the majority group, or the group which is socio-economically advantaged. Because of this the issue of bias in testing, including but not limited to DIF, has become an important political and legal concern.

While there is wide agreement as to the definition of DIF cited above, there is less agreement on how to identify DIF. Recently the Mantel-Haenszel (MH) statistic has emerged as one of the more widely used methods. Holland and Thayer (1988) introduced this statistic, developed in medical research, to the testing community. This procedure has gained popularity on both theoretical and practical grounds. Theoretically, it is consistent with the above definition of DIF, as it takes ability into account by blocking examinees on total test score. Within these "ability" groupings, the likelihood of passing an item is compared for the groups under study. This

---

[1]Paper presented at the meeting of NCME, Chicago, 1991.

[2]Laboratory of Psychometric and Evaluative Research Report No. 211. Amherst, MA: University of Massachusetts, School of Education.

approach has intuitive appeal for many practitioners. On the practical front, the Mantel-Haenszel procedure has economic advantages. First, it is relatively inexpensive to run in terms of computer time. Secondly, it requires far fewer examinees to yield meaningful results than IRT methods.

Research on the effect of sample size on detection rates of the MH procedure suggests that detection rates improve with larger samples. For instance, Wise (1987) conducted a simulation study which compared several bias detection methods, including the MH, at two different sample sizes, 400 and 800. He found the power of the MH increased substantially as sample size increased. He also noted that detection rates were best with items with moderately high discrimination (a=1.5) as opposed to items with lower (a=1.0) or higher (a=2.0) discrimination values. Rogers (1990) compared detection rates of the MH procedure with a logistic regression procedure. In this study she examined several variables which might be expected to have an impact on detection rates, including sample size. Looking at sample sizes of 250 and 500, she found, much like Wise, that increasing the sample size lead to a substantial increase in power.

While sample sizes of 500 and 250 are small by some standards, there are references in the literature to using the MH procedure with even smaller groups. For instance, Hills (1990) discusses the relative advantages and disadvantages of a number of DIF detection techniques. He suggests that the MH may be a good choice when sample sizes are between 100 and 300 for either or both groups.

While it has been generally accepted that the MH procedure will yield valid results with small sample sizes, the question of how small is small has not yet been answered. Therefore, the present study addresses the question of how many examinees are necessary to detect varying types and levels of

differential item functioning. Using simulated data, tests were constructed to contain differentially functioning items, where the item discrimination and difficulty parameters were known, and the latter differed for two groups by a known amount. It is possible to begin to determine how powerful this statistic is by analyzing these tests using the MH procedure at varying sample sizes.

Methods

Data for this study were generated using the program DATAGEN (Hambleton and Rovinelli, 1973). This program simulates examinee responses using the IRT model specified by the user. For this study a three parameter logistic model was used. Because our experience had suggested that the differences in underlying ability in the groups under study might be a significant variable, three data sets of 2000 examinees each were generated to allow for comparisons of groups with equal ability, and of groups where the focal group was less able. For two of these sets, the distributions of ability scores were set to a mean of zero. These will be referred to as Reference Group 1 and Focal Group 1. All distributions were normal with a standard deviation of one. These distributions were used to make comparisons between groups of equal ability. The abilities for the third distribution was set to a mean of -1.0. This group will be referred to as Focal Group 2. This was to allow for comparisons between groups with substantially different underlying ability distributions.

Five different 75 item tests were generated for each group. A 75-item test length was chosen as it is typical of many widely used standardized tests, such as achievement subtests, and is long enough to provide stable results. On each test the first 59 items were common items (the same across all five tests). IRT item statistics (b, a) for these 59 items were read in

using values selected randomly from published tables of statistics for the items for a recent administration of the Graduate Management Admission Test (Kingston, Leary, & Wightman, 1988). All c's for these 59 and all additional items were set to .20.

Eighty additional items were generated, with the parameters set to function differentially in the two groups. They were constructed so as to form 5 sets of 16 DIF items. Each set was combined separately with the 59 common non-DIF items to create 5 different 75 item tests for each of the groups described above. Each set of 16 items had four different levels of discrimination (a), with four items at each level. These values were read in to be .25, .60, .90, or 1.25 . To simulate DIF the values of the b parameters were set to differ by .25, .50, 1.00 or 1.50 for the reference and focal groups. The levels of a and the difference between the b's were completely crossed for each set of 16.

Five values of b were used for the reference group, the b's for the focal groups being increased by the amount specified above. These values were -2.5, -1.0, 0, 1.0, 2.5. Because there were five values of b, rather than four, it was not possible to completely cross each of these with the a's and b differences within each subtest. Therefore, four of these five values are represented within each subtest. However, within the entire set of 80 DIF items each level of b was completely crossed with each level of a and with each difference in b-values (i.e. the amount of DIF). Four sample item characteristic curve (ICC) comparisons are presented in Figure 1. ICC 1 and ICC 4 highlight the smallest and largest amounts, respectively, of bias simulated in the 80 items. ICC 2 and ICC 3 highlight middle levels of simulated bias in the reference and focal groups.

The MH procedure was run for each test comparison using a computer program written by H. Jane Rogers and the third author. Tests were first compared using all 2000 examinees in each group. The first 1000 examinees in each group were then selected, and the procedure was rerun. Because DATAGEN uses random generation procedures this was the same as randomly selecting examinees. This was repeated with 500, 200, and 100 examinees. In order to minimize the impact of chance variability, replications of the results were conducted at the smaller sample sizes. The 500 run was replicated once for each set, and the 200 and 100 runs were each replicated twice. While there were some inconsistencies in DIF identification across replications, in that some items were identified as DIF on one run but not on another, the overall pattern of results was very similar.

The results reported below are based on the second run of the MH program, using a .01 significance level. That is, items identified as DIF on the first run were removed from the calculation of the overall test score and then the MH statistic was recalculated for each of the 75 items in the test.

Results

A review of Table 1 reveals that the percentage of DIF items correctly identified as such decreased markedly as the number of examinees decreased. With 500 examinees in each group more than half of the DIF items were not flagged. With 1000 examinees, 58 and 61 percent of the DIF items were flagged (for the unequal and equal ability distributions respectively). This increased to 64 and 74 percent when the full sample of 2000 (each group) was used. The percentage of items correctly identified at sample sizes of 200 and 100 were very small. When the ability distributions for the groups were equal, the detection rates were consistently higher than when the

distributions were unequal, but, in both cases, the pattern across sample sizes was consistent.

All 80 DIF items are represented in Table 2, with items which the MH procedure detected as DIF indicated with an X in the appropriate row and column. There was a distinct pattern in the types of items which were flagged as DIF, and which were missed, and not surprisingly this pattern became more pronounced as sample size decreased.

The items which were missed at the 2000 examinee level were not identified at any other sample size with one exception. As the sample size decreased items were lost, but never gained. Poorly discriminating items were least likely to be identified, requiring larger sample sizes and greater differences between the two groups on item difficulty. The first items to be identified were those of moderate difficulty, with very difficult items being the least likely to be flagged. Not surprisingly, items with larger b differences were more likely to be identified than items with smaller b differences. Items most likely to be missed were the most difficult items, those with the smallest difference between the b's, and the most poorly discriminating items. This trend was apparent at the 2000 examinee level and became more marked as sample size decreased. The differences between the p-values for the equal ability distribution reference and focal groups are presented in Table 3, based on N=2000. By comparing Tables 2 and 3, it is possible to determine the pattern of p-value differences for items which were missed. When equal ability distributions are compared, the largest p-difference of a DIF item which was missed was .04, and the smallest p-difference of an item which was identified was .02 when groups of 2000 were used. With groups of 1000, these differences were .08 for the largest p-difference missed, and .03 for the smallest difference identified. WIth

groups of 500, a p-difference of .08 was missed, and .07 identified. With groups of 200, the largest p-difference missed was .17 and the smallest difference identified was .07. With groups of 100, the largest p-difference missed was .23, and the smallest identified was .15.

A similar pattern was apparent when groups of differing abilities were compared. (the p-differences reported here are based on a comparison of the two equal ability distributions, as comparing p-values for unequal ability distributions would not be meaningful.) In general, comparing groups of differing abilities resulted in larger differences being missed at all sample sizes. The largest p-differences missed were .07, .15, .17, .23, and .29 for groups of 2000, 1000, 500, 200, and 100 respectively. Conversely, with unequal ability distributions, the smallest p-differences of items identified were smaller than those identified with equal ability distributions. These differences were .01, .01, .03, .03, and .09 for groups of 2000, 1000, 500, 200, and 100 respectively. In general, these were associated with very easy items, so it is not surprising that they were more likely to be identified with unequal distributions.

Of the 59 non-DIF items, one was falsely identified as DIF fairly consistently at sample sizes of 1000 and 2000, with both equal and unequal ability distributions. A second item was consistently falsely identified with unequal ability distributions, at a sample size of 2000. A number of additional items were inconsistently flagged at the smaller sample sizes, but did not meet the criteria of being identified on at least two replications of the same set.

Discussion

The decrease in detection rates at the smaller sample sizes was not surprising. Any statistic will be less powerful as the sample becomes

smaller. However, the high percentage of items missed even at the largest sample sizes was unexpected. With 2000 examinees more than 30 percent of the DIF items were missed with unequal ability distributions, and more than 25 percent with equal distributions. While this is a relatively high percentage, an inspection of the p-value differences of the items missed when equal ability groups are compared reveals the differences are of little practical concern. Even if there were 10 items on a test with p-value differences of .03, this would be likely to result in less than half a point difference overall between the reference and focal groups. This is a level of DIF which most practitioners would probably find tolerable for most purposes. However, if ability distributions of the two groups are not equal, fairly substantial p-value differences can go undetected, even at this sample size.

Conversely, the amount of DIF missed when a sample size of 100 was used is more of a concern. Here it is likely that p differences of .20 would be missed routinely. Ten items with this amount of bias on a test could result in an overall difference between reference and focal groups of more than two points. A difference of this size could be a focus of concern, depending on the purpose of the test.

The implications for practitioners are clear. The results of the MH procedure are questionable at small sample sizes. The question of how small is small depends on the need for accuracy in identification. If only the most markedly DIF items are a concern, sample sizes of 200 in a group might be considered adequate. There would seem to be little justification for using sample sizes any smaller than 200. Sample sizes of 500 will yield more accurate results, and increasing to 1000 or 2000 will pick up all but the small p-value differences. Differences in ability distributions should be considered also, as comparisons of groups of differing abilities may impact

identification rates. Thus, if groups of differing abilities are to be compared, it is probably advisable to be even more conservative, and use large samples if possible. Practitioners should also be aware, however, that even with 2000 examinees per group with equal or unequal ability distributions some DIF items may not be identified. These are most likely to be very difficult items, poorly discriminating items, or items with relatively small differences in difficulty between the groups.

References

Hambleton, R. K., & Rovinelli, R. (1973). A FORTRAN IV Program for generating examinee response data from logistic test models. Behavioral Science, 18, 73-74.

Hills, J. (1990). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8, 5-11.

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity. Hillsdale NJ: Lawrence Erlbaum Associates.

Kingston, N., Leary, L., & Wightman, L. (1988). An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test. (GMAC Occasional Papers). Princeton, NJ: Graduate Management Admission Council.

Rogers, H. J. Item bias investigations with logistic regression. Unpublished doctoral dissertation, University of Massachusetts, 1990.

Wise, L. L. (1987, April). Differential item difficulty indicators in small samples. Paper presented at the meeting of the American Educational Research Association, Washington.

Table 1

Percentage of Differentially Functioning Items Correctly Identified

Sample Size/ Group	Reference and Focal Group Distributions Unequal	Equal
2000	64%	74%
1000	58%	61%
500	31%	38%
200	24%	28%
100	9%	18%

Note: The percentages reported for sample sizes of 500, 200 & 100 are based on the average number of items identified across replications.

Table 2

Items Identified as Differentially Functioning (Equal Ability Distributions)

Item Difficulty		Level of Item Discrimination									
		a=.25			a=.60		a=.90		a=1.25		
b[1]	b difference	Sample[2] 1 2 3 4 5			Sample 1 2 3 4 5		Sample 1 2 3 4 5		Sample 1 2 3 4 5		
-2.5	.25										X[3]
-1.0	.25							X X		X X	
0.0	.25					X		X X			X
1.0	.25					X X		X X			
2.5	.25										
-2.5	.50					X X		X X			X X
-1.0	.50					X X		X X X X			X X X
0.0	.50					X X		X X			X X
1.0	.50					X		X X			X
2.5	.50										
-2.5	1.00		X X X			X X		X X X			X X X X
-1.0	1.00		X X			X X X X		X X X X X			X X X X X
0.0	1.00		X X			X X X X		X X X X			X X X X
1.0	1.00		X X			X X X		X X X			X X X
2.5	1.00		X			X X		X			
-2.5	1.50		X X X			X X X X		X X X X X			X X X X
-1.0	1.50		X X X			X X X X X		X X X X X			X X X X X
0.0	1.50		X X X X			X X X X X		X X X X X			X X X X X
1.0	1.50		X X X X			X X X X X		X X X X X			X X X
2.5	1.50		X X			X					

[1]Item difficulty value for the reference group.

[2]Sample Size Per Group: 1 = 100 examinees; 2 = 200 examinees; 3 = 500 examinees; 4 = 1000 examinees; 5 = 2000 examinees.

[3]X indicates that the item was identified as DIF. (For sample sizes where replications were run, an X indicates identification on at least two runs.)

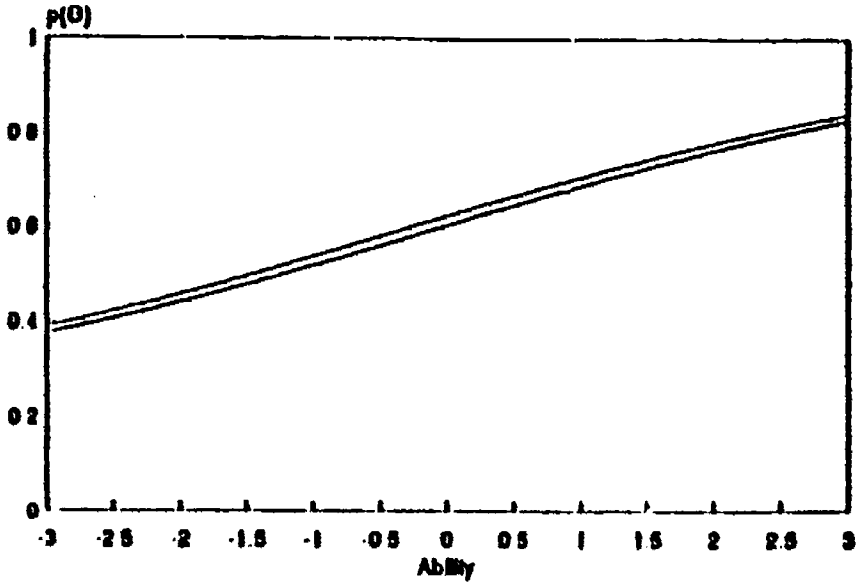
Table 3

P Value Differences for DIF Items (Equal Ability Distributions, N=2000)

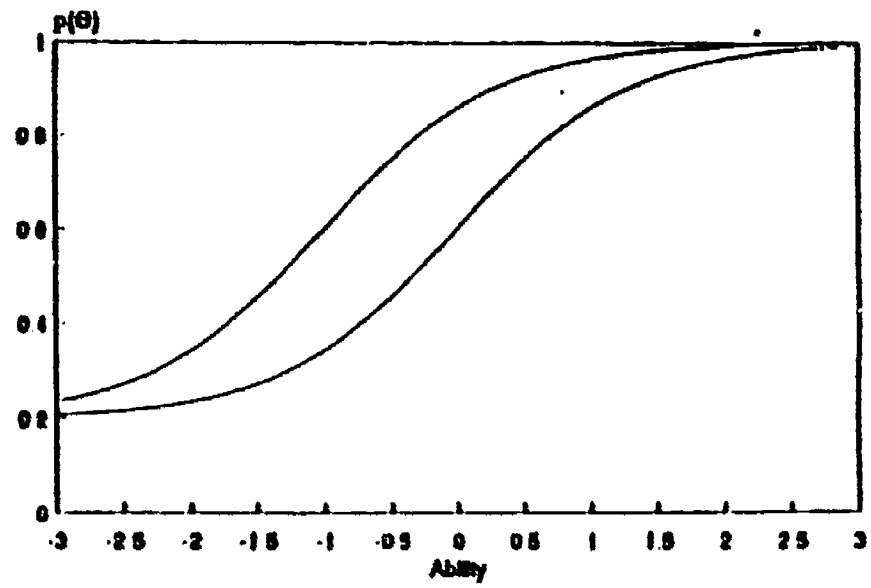
Item Difficulty		Level of Item Discrimination			
b-value[1]	b-value Difference	.25	.60	.90	1.25
-2.5	.25	.02	.01	.01	.02
-1.0	.25	.01	.03	.05	.05
0.0	.25	.02	.04	.08	.06
1.0	.25	.00	.06	.07	.04
2.5	.25	.04	.01	.00	.01
-2.5	.50	.01	.06	.03	.03
-1.0	.50	.03	.06	.12	.09
0.0	.50	.03	.09	.10	.11
1.0	.50	.03	.06	.10	.07
2.5	.50	.01	.03	.01	.02
-2.5	1.00	.09	.08	.07	.07
-1.0	1.00	.10	.16	.20	.23
0.0	1.00	.08	.16	.19	.23
1.0	1.00	.07	.12	.12	.13
2.5	1.00	.04	.06	.04	.02
-2.5	1.50	.11	.17	.16	.15
-1.0	1.50	.13	.22	.31	.34
0.0	1.50	.15	.22	.24	.29
1.0	1.50	.14	.15	.15	.17
2.5	1.50	.10	.06	.02	.03

[1]Item difficulty value for the reference group.

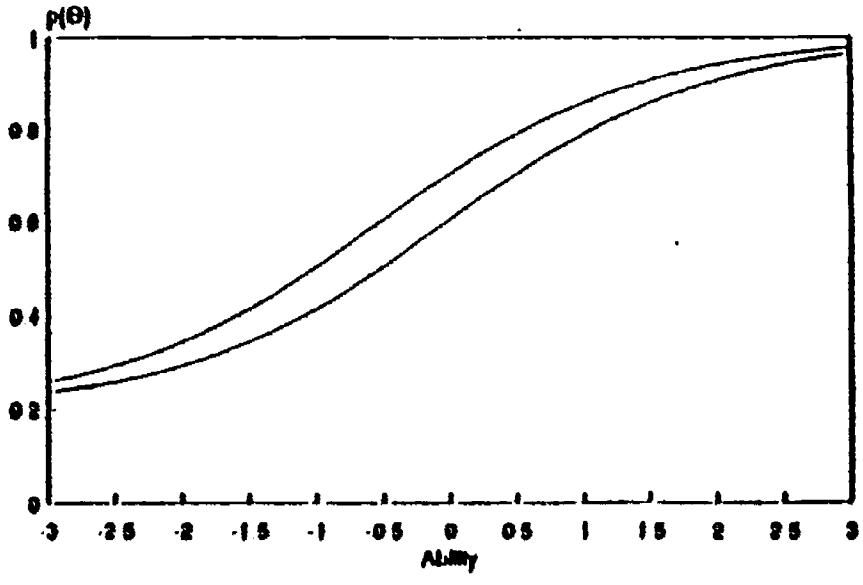
ICC 1
a = .25, b difference = .25



ICC 3
a = .90, b difference = 1.00



ICC 2
a = .60, b difference =.50



ICC 4
a = 1.25, b difference = 1.50

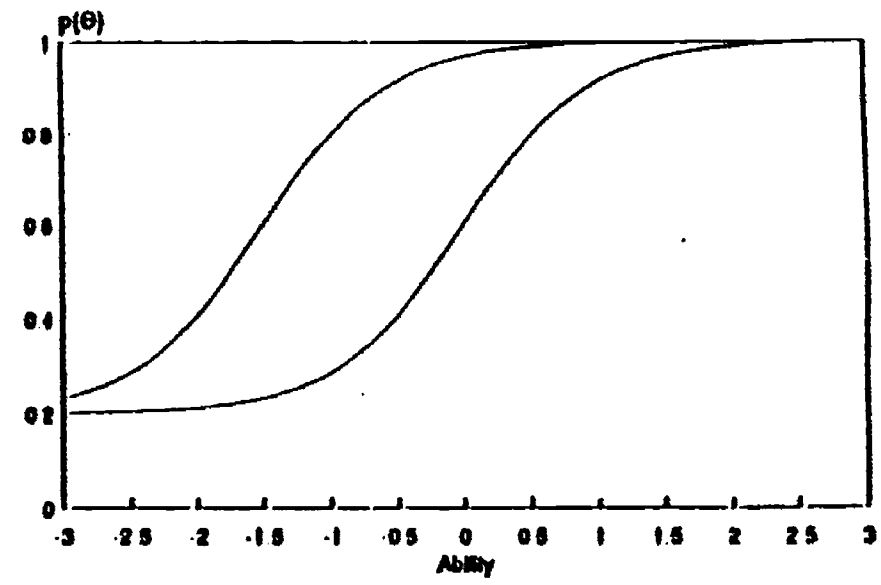


Figure 1. A sample of the biased items in the computer simulation study.